

Analytical Study of Key Techniques for Cross-Modal Feature Alignment and Decision-Level Fusion in Brain–Computer Interface–Virtual Reality Systems [†]

Dan Liu 

Faculty of Digital Creativity, Changzhou College of Information Technology, Changzhou 213000, China; dan957329@gmail.com

[†] Presented at the 9th Eurasian Conference on Educational Innovation 2026 (ECEI 2026), Da Nang City, Vietnam, 30 January–2 February 2026.

Abstract

Feature alignment and decision-level fusion in multimodal BCI–VR interaction were investigated using Transformer-based cross-modal embeddings, Lab Streaming Layer time synchronization, attention masks, and wavelet filtering for robust representation. A four-modal acquisition and synchronization platform covering electroencephalography, electromyography, eye-tracking, and speech was constructed, and fusion was achieved by introducing a stacking meta-learner together with a confidence-aware dynamic weighting mechanism. Prototype validation and comparative evaluations were conducted on virtual reality (VR) target-selection, trajectory-following, and object-manipulation tasks. The results showed that the proposed approach outperformed baselines such as weighted voting and independent single-modality classifiers in accuracy, cross-session and cross-subject generalization, and noise robustness, while achieving a measurable reduction in end-to-end response latency, indicating that an integrated semantic alignment–adaptive fusion pipeline enhanced stable outputs and robustness in multimodal interaction. The unified semantic alignment model tailored to BCI–VR can be used for establishing an integrated engineering workflow spanning synchronization, robust representation, and adaptive fusion, and for providing transferable evaluation metrics and application paradigms that offer methodological and technical references for scenarios such as rehabilitation training, virtual education, and intelligent control.

Keywords: brain–computer interface (BCI); virtual reality (VR); cross-modal feature alignment; decision-level fusion; time synchronization

1. Introduction

Brain–computer interface (BCI)-based virtual reality (VR) systems comprise a BCI reception/control subsystem and a three-dimensional VR rendering subsystem [1]. This architecture transforms traditional physical interfaces toward multimodal intelligent interaction. The incorporation of BCI elevates human–machine integration beyond what can be achieved with VR devices alone [2]. However, pronounced heterogeneity across modalities—in physical properties, temporal characteristics, and semantic representations—makes realizing effective coordination, precise synchronization, and reliable fusion the core technical bottleneck that limits BCI–VR performance.

In real-world settings, three pain points are prominent: (1) distribution shift induced by cross-session and cross-subject variation undermines model stability and generaliza-



Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 29 June 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

tion; (2) acquisition asynchrony and jitter accumulate across modalities, weakening event alignment and temporal causality; and (3) commonly used decision-level fusion schemes rely on static or heuristic weighting, are insensitive to noise and artifacts, and tend to raise end-to-end latency when robustness is pursued. Without jointly addressing these issues, the advantages of multimodality are difficult to translate into repeatable performance gains on VR tasks such as target selection, trajectory following, and object manipulation.

To tackle these challenges, a unified semantic alignment model tailored to BCI–VR was developed in this study to leverage Transformer-based cross-modal embeddings with a contrastive alignment loss to mitigate semantic bias under modality heterogeneity and domain shift. An integrated engineering workflow was established to span acquisition–synchronization–robust representation–adaptive fusion, providing reusable implementation paradigms for event triggering, clock correction, masking strategies, and stacking-based fusion with confidence-aware dynamic weighting. Transferable evaluation metrics and application paradigms were used to validate the approach on three VR tasks, demonstrating improvements in accuracy, cross-session and cross-subject generalization, and noise robustness, together with a quantifiable reduction in end-to-end latency, offering methodological and technical references for scenarios such as rehabilitation training, virtual education, and intelligent control.

2. Literature Review

Existing multimodal BCI–VR studies have tried to address cross-modal semantic alignment, time synchronization, and decision-level fusion as separate modules. However, it is difficult on the engineering side to form a closed loop spanning synchronization–representation–alignment–fusion, and such pipeline fragmentation makes it hard to simultaneously ensure accuracy, generalization, and low latency.

Regarding cross-modal alignment, Vafaei et al. review that Transformers leverage self-attention to explicitly model heterogeneous dependencies within a shared semantic space and have become a key direction for aligning signals such as EEG [3]. However, their practical effectiveness strongly depends on synchronization quality and masking strategies for missing segments. Wan et al. propose EEGformer, which uses multi-stage Transformers to jointly capture the spatiotemporal structure of EEG, indirectly supporting the feasibility of a unified semantic space while highlighting the decisive impact of upstream engineering quality [4]. Kothe et al. systematically describe Lab Streaming Layer (LSL), showing that, under multi-device setups, LSL provides sampling-level timestamping with drift compensation and jitter management and has effectively become a de facto standard for neuro-behavioral multimodal synchronization [5].

Nevertheless, most work emphasizes mean error while under-reporting jitter distributions (P95/P99), late packets, and event-window tolerance, thereby constraining reproducibility and cross-platform stability. Vortmann et al. compare early/mid/late fusion of EEG and eye tracking, finding that late fusion performs better for attention classification but struggles to re-weight modalities dynamically according to real-time data quality [6]. Younis et al. conduct a systematic evaluation of ensemble/stacking methods on multimodal physiological signals, showing that stacking yields more robust gains under inconsistent evidence [7]. Multiple studies indicate that combining electroencephalography (EEG) with eye tracking, speech, and electromyography (EMG) can enhance recognition and responsiveness in immersive tasks; however, robustness gains are jointly bounded by synchronization quality, handling of missing segments, and denoising strategies.

Based on these findings, the core concepts were defined in this study as follows. Cross-modal semantic alignment refers to learning geometric proximity for the same interaction intent within a shared latent space via modality-specific encoders and a cross-modal

Transformer trained with a task loss plus a contrastive alignment loss, while attention masks exclude artifact-contaminated or missing segments from alignment computation. Decision-level fusion denotes a late-fusion scheme of calibrated modality posteriors → stacking meta-learner, where sample-wise weights are determined dynamically by confidence, short-term stability, and noise estimates. Robust representation is constrained to a wavelet-based denoising + attention-masking combination that is applied consistently across the alignment-to-fusion pipeline. Time synchronization is based on LSL, with reporting that includes jitter distributions (P95/P99), late-packet rates, and event-window tolerance, and these are mapped to alignment error, weight volatility, and end-to-end latency. End-to-end latency focuses on net change rather than single-point overheads. Generalization and robustness are evaluated using fixed protocols: leave-one-session-out (cross-session), leave-one-subject-out (LOSO), and controlled perturbation-strength curves; ablation studies disentangle the marginal contributions of alignment, masking, and dynamic weighting.

3. Methodology

The proposed framework centers on a semantic alignment–adaptive fusion pipeline, aiming to stably improve the accuracy and transferability of multimodal interaction under controllable latency constraints. Two design principles are adopted: (1) treat time synchronization and data availability as prerequisites that influence model decisions, and feed their quantitative indicators (including jitter distributions and late-packet ratios) directly into subsequent modeling and interpretation; and (2) embed robustness as an internal mechanism rather than a post hoc remedy, maintaining a consistent visibility criterion across the entire pipeline via wavelet denoising and attention masks. The resulting system pathway is: synchronization/alignment → robust representation → semantic alignment → decision fusion → VR feedback. All stages operate on a unified data timeline to reduce decision bias caused by information fragmentation.

3.1. Four-Modal Platform and Data Flow

The system comprises four signal sources: EEG, EMG, eye tracking, and speech. To unify temporal granularity and processing cadence, EEG is acquired at 512 Hz with 32 channels using a 200 ms analysis window and a 50 ms hop; EMG is acquired at 1 kHz with 8 channels using 100 ms/25 ms; eye tracking streams binocular data at 120 Hz in real time (8.3 ms granularity, no overlap); and speech is sampled at 16 kHz with 25 ms frames and a 10 ms hop. The upper bounds for available data arrival latency $T_{\text{acq}}^{\text{max}}$ and first-layer encoding latency $T_{\text{enc}}^{\text{max}}$ are constrained, respectively, to 25 ms/35 ms for EEG, 15 ms/25 ms for EMG, 12 ms/10 ms for eye tracking, and 18 ms/22 ms for speech, corresponding to single-modality upper T_m^{max} limits of 60, 40, 22, and 40 ms (see Table 1). Data flow follows a publish–subscribe model with ring buffers; when the downstream is briefly blocked, cached segments far from the event are discarded first while preserving windows near the event, ensuring that subsequent early decisions are not hindered by queue backlogs. On the VR side, event codes are injected at key moments—target presentation, trajectory start/stop, and object grasp—to serve as anchors for alignment and supervision.

Table 1. Four-modal sampling configuration and end-to-end latency upper bounds.

Modality	Sampling Rate (Hz)	Number of Channels	Quantization Bit Depth (bits)	Expected Band-Pass/Processing	Framing-Hop (Analysis Window)	T_{acq}^{max}	T_{enc}^{max}	T_m^{max}
EEG	512 Hz	32	24-bit	0.5–40 Hz + Power-line notch	200 ms Frame/ 50 ms hop	25 ms	35 ms	60 ms
EMG	1 kHz	8	16-bit	20–450 Hz	100 ms frame/ 25 ms hop	15 ms	25 ms	40 ms
Eye tracking	120 Hz	2 (Binocular)	16-bit	Velocity threshold + drift correction	Real-time/no overlap (~8.3 ms temporal granularity)	12 ms	10 ms	22 ms
Speech	16 kHz	1	16-bit	VAD + 20 ms STFT	25 ms frame/ 10 ms hop	18 ms	22 ms	40 ms

3.2. Time Synchronization and Event Triggering

Cross-device clocks are aligned and drift-compensated using LSL timestamps. The source clock t_s and the local clock t_l approximately satisfy a linear relationship $t_l \approx at_s + b$ within a sliding window. A least-squares estimate $\hat{a}, \hat{b}$, is used to apply a unified time correction to newly arrived data as

$$\tilde{t} = \hat{a} t_s + \hat{b} \quad (1)$$

In Equation (1), t_s denotes the timestamp generated by the source device, t_l denotes the local reference timestamp, W denotes the sliding calibration window, the estimated clock skew and the estimated clock offset are obtained by least squares, and $\tilde{t}$ denotes the corrected timestamp used for subsequent multimodal alignment.

To handle network jitter and late packets, modality-specific tolerance windows $[-\tau_-, +\tau_+]$ are set as follows: EEG $[-30, +50]$ ms, EMG $[-25, +40]$ ms, eye tracking $[-35, +55]$ ms, and speech $[-20, +35]$ ms (Table 2). Samples arriving within the window are aligned to events; those exceeding the upper bound are labeled invisible. Synchronization quality reporting goes beyond mean bias μ_{offset} and includes jitter P95/P99 and late-packet ratio ρ_{late} , as well as the interpolation rate ρ_{interp} . Under the default acceptance criteria in this study, residual mean bias ≤ 1.0 ms, P95 ≤ 5.0 ms, P99 ≤ 8.0 ms, and $\rho_{late} \leq 0.35\%$. End-to-end latency is decomposed as

$$T_{e2e} = T_{acq} + T_{sync} + T_{\downarrow} + T_{align} + T_{fuse} + T_{vr}, \quad (2)$$

where T_{sync} is conservatively estimated using the upper bounds of residual bias and jitter. In practice, $\mu_{offset} + P95$ (e.g., ≈ 5.3 ms) is substituted to ensure consistency between statistical interpretation and engineering implementation.

Table 2. Statistical criteria for synchronization error and jitter.

Modality	μ_{offset}	Jitter P95 (ms)	Jitter P99 (ms)	ρ_{late}	ρ_{interp}	$[-\tau_-, +\tau_+]$
EEG	0.8	4.0	6.0	0.30	0.20	($[-30, +50]$)
EMG	0.6	3.0	5.0	0.20	0.15	($[-25, +40]$)
Eye Tracking	1.2	8.0	12.0	0.40	0.30	($[-35, +55]$)
Speech	0.9	3.5	5.5	0.25	0.20	($[-20, +35]$)
Merged (after compensation)	≤ 1.0	≤ 5.0	≤ 8.0	≤ 0.35	≤ 0.25	Use the wider one per modality

In Equation (2), T_{e2e} denotes the end-to-end latency of one interaction cycle; T_{acq} denotes data-acquisition and availability waiting time; T_{sync} denotes synchronization time;

$T_{\downarrow}$ denotes low-level or first-layer processing time; T_{align} denotes semantic-alignment time; T_{fuse} denotes decision-fusion time; and T_{vr} denotes VR rendering and feedback time.

3.3. Robust Representation: Wavelet Denoising and Attention Masks

Physiological signals exhibit pronounced non-stationary noise and artifacts. For EEG/EMG, discrete wavelet decomposition (Daubechies-4 as the mother wavelet) is applied with decomposition depth adapted to the sampling rate, followed by channel-wise soft-threshold denoising and reconstruction. For eye tracking, blink events are localized via a velocity threshold and corrected using polynomial baseline fitting. For speech, endpoint detection (VAD) combines energy and spectral entropy, with suppression of short silences and plosive bursts.

Departing from the traditional preprocess-and-stop paradigm, we explicitly encode data availability as attention masks: binary/probabilistic masks jointly generated from synchronization assessment (late packets, interpolated segments), VAD, and artifact detection. These masks are propagated through subsequent self-attention computations, masking unusable segments so that they neither contribute to similarity measures nor propagate gradients downstream. Masks are kept consistent between training and inference to avoid train-deploy specification drift. This treatment maps directly to the criteria in Table 2: any late packet outside the tolerance window is set to 0 uniformly, and very short missing segments are not time-interpolated to prevent spurious associations from misleading the alignment targets.

3.4. Cross-Modal Semantic Alignment

Frame-level feature sequences $\{x_t^{(m)}\}$ are obtained from modality-specific encoders. To ensure comparable capacity, EEG/EMG use lightweight temporal convolutions or a temporal Transformer, while eye tracking and speech use compact time-/frequency-domain encoders. The outputs are then concatenated into a token sequence within a shared cross-modal Transformer, augmented with modality identifiers and temporal positional encodings, and the visibility masks from the previous section are applied. The model is optimized with a joint objective.

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{align}}, \quad (3)$$

where $\mathcal{L}_{\text{task}}$ uses cross-entropy for discrete intent tasks and Huber regression loss for continuous control tasks; $\mathcal{L}_{\text{align}}$ adopts an InfoNCE-style formulation, pulling together cross-modal embeddings that correspond to the same interaction intent within the same event window and pushing apart those of different intents.

In Equation (3), $\mathcal{L}$ denotes the overall training objective, $\mathcal{L}_{\text{task}}$ denotes the task-specific loss, $\mathcal{L}_{\text{align}}$ denotes the cross-modal alignment loss, and λ_{align} denotes the weighting coefficient used to balance alignment against task optimization.

$$\mathcal{L}_{\text{align}} = -\sum_i \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+) / \tau)}{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+) / \tau) + \sum_{j \neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j^-) / \tau)} \quad (4)$$

Here, in Equation (4), sim denotes cosine similarity under temperature τ , $\mathbf{z}_i^+$ are same-intent embeddings from paired modalities, and $\mathbf{z}_j^-$ are negative samples. Because event-level alignment is bounded by the jitter and late-packet constraints in Table 2, positive/negative sampling for $\mathcal{L}_{\text{align}}$ is strictly cropped by the tolerance window to avoid false positives caused by temporal mismatch.

3.5. Decision-Level Fusion: Stacking Meta-Learning with Confidence-Aware Dynamic Weighting

Cross-modal embeddings yield posterior distributions or continuous estimates on each modality branch. To obtain comparable probability scales, temperature scaling to single-modality outputs is applied for discrete tasks: given uncalibrated logits z , the calibrated probability is $\sigma(z/T)$, with temperature $T > 0$ fitted on the validation set and evaluated using ECE/Brier metrics. The second-stage fusion adopts a stacking meta-learner (either linear or a small tree model), whose inputs comprise calibrated posteriors from each modality + quality indicators. To maintain stable outputs when signal quality fluctuates and evidence conflicts, sample-level dynamic weights are introduced.

$$w_m^{(t)} = \frac{\exp(\alpha c_m^{(t)} + \beta s_m^{(t)} - \gamma n_m^{(t)})}{\sum_k \exp(\alpha c_k^{(t)} + \beta s_k^{(t)} - \gamma n_k^{(t)})}, \quad (5)$$

In Equation (5), $w_m^{(t)}$ denotes the dynamic weight assigned to modality m , $c_m^{(t)}$ denotes calibrated confidence (e.g., a monotonic transform of maximum posterior or log-likelihood), $s_m^{(t)}$ denotes short-term stability (the negative of posterior variance/entropy within a sliding window), $n_m^{(t)}$ denotes the noise estimate for modality m , and alpha, beta, and gamma denote the weighting coefficients that control the relative contributions of confidence, stability, and noise, respectively (derived from artifact detection and energy ratios).

For real-time policy, based on the hop in Table 1 and per-modality T_m^{max} , we set a maximum waiting threshold Δ_{max} (50 ms in this study). If a key modality update is not received within Δ_{max} , an early decision is triggered to prevent latency increases caused by waiting for agreement. Latency is reported using the mean and P95, and tail risks for T_{sync} are interpreted in conjunction with the jitter statistics in Table 2.

3.6. Reproducibility Essentials and Parameter Specifications

To avoid capacity confounds, the number of layers and heads in the cross-modal Transformer is kept in the same order of magnitude as the total parameter count of the single-modality encoders; $\lambda \in [0.1, 0.5]$, $T \in [1.0, 3.0]$, $\alpha, \beta, \gamma \in [0, 2]$ are selected via coarse grid search; the same mask-generation strategy and the same event-window length are used for both training and inference. Cross-session and cross-subject evaluations uniformly adopt LOSO; noise robustness is assessed by injecting artifacts/drop frames/noise of controlled intensity to obtain performance–intensity curves. End-to-end evaluation strictly follows Equation (2) with sub-components recorded.

T_{acq} and T_{enc} use the default upper bounds in Table 1 as budgets; T_{sync} is estimated as $\mu_{offset} + P95$; T_{align} and T_{fuse} come from module-level timing; T_{vr} is obtained from engine-side rendering-frame statistics. If any session's ρ_{late} or P99 exceeds the merged acceptance thresholds in Table 2, it is excluded from the primary analysis while retained in sensitivity analyses, ensuring that conclusions are not confounded by synchronization quality.

4. Experiment

The objective of the experimental design is to evaluate the effectiveness and transferability of the proposed semantic alignment–adaptive fusion framework under controllable end-to-end latency constraints across three representative VR task types. A total of 24 healthy adult participants (age 20–34 years; 12 male, 12 female; 22 right-handed) were recruited. All participants provided informed consent and passed safety screening for human–computer interaction tasks. To control for learning effects, a 15 min familiarization session was administered before the formal experiment; this warm-up data was not included in the analysis. Each participant completed two sessions (S1 and S2) over two days,

separated by 48 ± 12 h, in order to induce cross-session distributional variation. In each session, participants performed three tasks in sequence: (A) target selection, (B) trajectory following, and (C) object manipulation. Task order was counterbalanced across participants using a Latin square design.

1. Target selection: An 8-target radial layout was used. Each trial consisted of 1 s gaze calibration, 2 s target presentation, a 3 s selection window, and 1 s feedback. Participants completed selection using a coordinated strategy combining gaze fixation + short spoken command + subtle muscular activation/motor imagery. Each session contained 240 trials ($8 \text{ targets} \times 30 \text{ repetitions}$), for a total of 480 trials per participant;
2. Trajectory following: The participants were instructed to continuously follow a moving control point in 3D space for 20 s using combined gaze + speech rhythm + EMG/EEG intent. The trajectory included segments of straight motion, spline paths, and perturbed segments. Each session contained 12 segments (including 4 perturbed trajectories), for a total of 24 segments per participant;
3. Object manipulation: This task comprised three phases: grasp, move, and place. The mass and position of the target object were randomized across three difficulty levels. Each session contained 60 trials ($\text{difficulty} \times \text{position} \times \text{repetitions}$), for a total of 120 trials per participant.

Before each session, a 90 s resting + synchronization calibration phase was conducted to verify that residual bias under LSL, jitter P95/P99, and late-packet ratio satisfied the merged acceptance thresholds defined in Table 2. Sessions that failed these criteria were immediately restarted and recalibrated. During the formal tasks, packets exceeding the event-window tolerance were marked as invisible segments according to the defined rules, and corresponding attention masks were generated to suppress their influence in both alignment and fusion stages.

To evaluate robustness, controlled perturbations of varying intensity were applied to three modality classes under safety and ethics constraints.

1. For EEG/EMG, controlled blink/micro-movement induction segments were inserted between task blocks, and adjacent segments were filtered in analysis based on an energy threshold;
2. For speech, contextual noise was added with signal-to-noise ratios of 20/10/0 dB;
3. For eye tracking, pseudo frame drops were introduced at 5/10/20%, with drop patterns sampled from the empirical late-packet distribution.

Perturbed and clean segments were both included in training. However, at evaluation time, performance was reported separately by perturbation level to obtain performance–intensity curves. Extremely short missing segments were not time-interpolated and were only handled via masking, in order to avoid introducing spurious alignment.

Without exclusion, per participant across both sessions, the annotated events were approximately: 480 trials for target selection; 24 trajectory segments (total duration ≈ 480 s) for trajectory following; and 120 trials for object manipulation. Across all participants, this corresponds to roughly 11,520 discrete trials and 192 min of continuous segments (excluding warm-up). Two partitioning protocols were used as follows.

1. Cross-session: for each participant, S1 was used for training and S2 for testing, and vice versa, to obtain bidirectional validation;
2. Cross-subject: LOSO, training on 23 participants and testing on the held-out participant in rotation.

These protocols respectively probe distributional shift caused by within-subject temporal variation and by inter-individual differences. To avoid capacity confounds, all training

baselines used the same number of optimization epochs and identical early-stopping criteria; the validation set was used only for temperature scaling and for selecting dynamic-weight parameters.

Under this design, the three tasks jointly cover three interaction paradigms—discrete intent, continuous control, and composite manipulation. The two partitioning protocols separately target session-level non-stationarity and subject-level variability. Controlled perturbations affect both physiological and behavioral channels. This enables observed changes in accuracy, generalization, robustness, and end-to-end latency to be directly attributed to synchronization quality, masking strategy, and dynamic fusion mechanisms, rather than to external engineering artifacts.

5. Results and Discussions

5.1. Results

Across all three tasks, the proposed method outperformed the baselines in both accuracy and latency. For target selection, the overall accuracy and macro-averaged F1 (pooled over cross-session/cross-subject) reached 91.8 and 91.2%, respectively. Compared with Best Single Modality (86.2/85.7%), Static Weighted Voting (88.1/87.5%), and Late Averaging without Alignment (89.0/88.6%), the gains were +5.6/+3.7 percentage points. For object manipulation, measured by task success rate and macro F1, the proposed method achieved 88.5/87.9%, improving over the three baselines by +7.1/+4.4 percentage points.

For trajectory following, evaluated by trajectory error (root mean square error (RMSE), m) and normalized path deviation, the proposed method yielded 0.126 m/0.108, which is lower than Best Single Modality (0.163/0.142), Static Weighted Voting (0.151/0.131), and Late Averaging without Alignment (0.147/0.127) by 22.7/24.0%, 16.6/17.6%, and 14.3/15.0%, respectively. A repeated-measures ANOVA indicated a significant main effect of method; Holm-corrected post hoc comparisons showed that the proposed method was significantly better than each of the three baselines, with effect sizes Cohen's $d = 0.62 \sim 0.94$. For end-to-end latency, the proposed method's mean/P95 were 128 ms/182 ms, representing reductions of 11.1/12.1% versus Static Weighted Voting (144/207 ms) and 7.9/8.1% versus Late Averaging without Alignment (139/198 ms); differences were significant.

5.2. Generalization (Cross-Session/Cross-Subject)

Under the cross-session protocol, accuracy for target selection and object manipulation improved by +6.3/+5.1 pp relative to Late Averaging without Alignment; trajectory RMSE was further reduced by 13.2%. Under cross-subject (LOSO), the proposed method improved over Static Weighted Voting by +5.5 pp (target selection) and +4.7 pp (object manipulation), with trajectory RMSE reduced by 11.6%. Improvements were consistent across tasks for both protocols (non-significant interaction), suggesting that a unified semantic space provides stable benefits against subject variability and session drift.

5.3. Noise Robustness

In controlled-perturbation tests (speech signal to noise ratio (SNR) = 20/10/0 dB; eye-tracking pseudo frame-drop 5/10/20%; EEG/EMG artifact segments filtered by energy threshold), the proposed method exhibited shallower degradation slopes in the performance–intensity curves across tasks. For target selection, at SNR = 0 dB, the accuracy drop was 6.9%, substantially better than Static Weighted Voting (14.8%) and Late Averaging without Alignment (12.5%). With 20% eye-tracking frame-drop, accuracy decreased by 7.4%, whereas the baselines dropped by 15.1% and 12.9%, respectively. For trajectory following, at SNR = 0 dB, RMSE increased by only +0.026 m, compared with +0.044/+0.038 m for the two baselines.

5.4. Latency Analysis

Recorded by components [Equation (2)], the proposed method introduces an additional +7.4 ms on average (6.8–8.1 ms across tasks) relative to Static Weighted Voting. However, due to the early decision policy with a maximum waiting threshold of =50 ms, the decision-wait time decreased by 23.1 ms on average (P95 reduced by 28.4 ms), yielding a net reduction overall. Session-to-session fluctuations on the synchronization side (estimated by mean bias + P95) are moderately correlated with the P95 of (), but within accepted sessions, the slope of this relationship is insufficient to offset the gains from early decision. Module-level timing further shows that when the eye-tracking or speech channel experiences high-intensity perturbations, the dynamic weights quickly down-regulate their contributions, avoiding ineffective delays from waiting for agreement, and thereby maintaining a positive net benefit. Table 3 summarizes the main ablation results, where the omission of semantic alignment, dynamic weighting, or wavelet denoising consistently reduced classification accuracy and F1, increased trajectory RMSE, and tended to increase P95 latency.

Table 3. Key ablations and significance tests (mean $\pm$ standard deviation).

Variant	Target-Selection Accuracy (%)	Δ pp	p	(d)	F1-Score (%)	Δ pp	p	(d)	RMSE (m)	Δ pp	(p)	(d)	P95 (ms)	Δ pp	p	(d)
Proposed	91.8 $\pm$ 3.1	—	—	—	87.9 $\pm$ 3.6	—	—	—	0.126 $\pm$ 0.018	—	—	—	182 $\pm$ 21	—	—	—
Without align	87.2 $\pm$ 3.5	−4.6	<0.01	0.72	82.7 $\pm$ 3.9	−5.2	<0.01	0.76	0.145 $\pm$ 0.020	+15.1%	<0.01	0.73	191 $\pm$ 24	+9 ms	0.018	0.41
Without dynamic	89.2 $\pm$ 3.3	−2.6	0.012	0.58	85.8 $\pm$ 3.7	−2.1	0.021	0.55	0.136 $\pm$ 0.019	+8.4%	0.024	0.52	193 $\pm$ 23	+11 ms	<0.01	0.60
Without wavelet	90.0 $\pm$ 3.2	−1.8	0.034	0.44	86.1 $\pm$ 3.5	−1.8	0.041	0.43	0.133 $\pm$ 0.019	+5.9%	0.029	0.44	185 $\pm$ 22	+3 ms	0.11	0.20

Note: Δ pp denotes the absolute change in percentage points relative to the full proposed model for the percentage-based metrics. For RMSE, Δ denotes the relative percentage increase compared with the proposed model, whereas for P95, it denotes the absolute increase in milliseconds. p denotes the p -value of the post hoc comparison against the proposed model, and d denotes Cohen's d effect size.

6. Conclusions

Guided by a semantic alignment–adaptive fusion paradigm, an end-to-end framework encompassing synchronization, robust representation, alignment, and fusion is presented in this study. The proposed framework demonstrates consistent improvements in accuracy, cross-session/cross-subject generalization, and robustness under noise conditions across three VR tasks, while achieving a quantifiable reduction in end-to-end latency.

Considering the results and limitations of this study, two complementary trajectories are required. First, the current multimodal suite (EEG, EMG, eye tracking, and speech) will be augmented with additional signals, such as haptics, acoustic myography, and environmental semantics. By leveraging cross-modal knowledge distillation and structured priors, this path aims to maintain a sparse, interpretable alignment space while compressing model capacity to mitigate training and deployment constraints. Second, to address session drift and device heterogeneity in practical applications, lightweight online domain adaptation must be integrated with cache distillation. This approach combines operator fusion, quantization, and early-exit strategies to ensure stable, real-time inference on low-compute platforms. The overarching objective of both trajectories is to further enhance model transferability and deployability without compromising latency or the reproducible empirical criteria established in this work.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Changzhou College of Information Technology (Ethical review and approval were waived).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Teng, J.; Huang, J.-H.; Gong, K. Design of virtual reality simulated driving teaching system based on BCI. *J. Graph.* **2020**, *41*, 217–223.
2. Wu, X. Is virtual reality under brain–computer interface an illusion? Reflection on Chalmers’ simulation hypothesis. *Stud. Dialect. Nat.* **2023**, *39*, 76–82. [[CrossRef](#)]
3. Vafaei, E.; Hosseini, M. Transformers in EEG analysis: A review of architectures and applications in motor imagery, seizure, and emotion classification. *Sensors* **2025**, *25*, 1293. [[CrossRef](#)] [[PubMed](#)]
4. Wan, Z.; Li, M.; Liu, S.; Huang, J.; Tan, H.; Duan, W. EEGformer: A transformer-based brain activity classification method using EEG signal. *Front. Neurosci.* **2023**, *17*, 1148855. [[CrossRef](#)] [[PubMed](#)]
5. Kothe, C.; Shirazi, S.Y.; Stenner, T.; Medine, D.; Boulay, C.; Grivich, M.I.; Artoni, F.; Mullen, T.; Delorme, A.; Makeig, S. The Lab Streaming Layer for synchronized multimodal recording. *Imaging Neurosci.* **2025**, *3*, IMAG.a.136. [[CrossRef](#)] [[PubMed](#)]
6. Vortmann, L.-M.; Ceh, S.; Putze, F. Multimodal EEG and eye tracking feature fusion approaches for attention classification in hybrid BCIs. *Front. Comput. Sci.* **2022**, *4*, 780580. [[CrossRef](#)]
7. Younis, E.M.G.; Zaki, S.M.; Kanjo, E.; Houssein, E.H. Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. *Sensors* **2022**, *22*, 5611. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.